RECEIVED
CENTRAL FAX CENTER

MAY 3 0 2006

Yee & Associates, P.C.

4100 Alpha Road
Suite 1100
Dallas, Texas 75244

Main No. (972) 385-8777
Facsimile (972) 385-7766

Facsimile Cover Sheet

To: Commissioner for Patents for **Examiner Huyen X. Vo** Group Art Unit 2655	Facsimile No.: **571/273-8300**
From: Candace Crawford Legal Assistant to Peter Manzo	No. of Pages Including Cover Sheet: **28**

Message:

Enclosed herewith:

- Transmittal of Appeal Brief; and
- Appeal Brief.

Re: Application No. **09/920,983**
Attorney Docket No: **DE920000060US1**

Date: Tuesday, May 30, 2006

Please contact us at (972) 385-8777 if you do not receive all pages indicated above or experience any difficulty in receiving this facsimile.

This Facsimile is intended only for the use of the addressee and, if the addressee is a client or their agent, contains privileged and confidential information. If you are not the intended recipient of this facsimile, you have received this facsimile inadvertently and in error. Any review, dissemination, distribution, or copying is strictly prohibited. If you received this facsimile in error, please notify us by telephone and return the facsimile to us immediately.

PLEASE CONFIRM RECEIPT OF THIS TRANSMISSION BY FAXING A CONFIRMATION TO 972-385-7766.

RECEIVED
CENTRAL FAX CENTER
MAY 30 2006

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of: **Haase et al.** §
Serial No.: **09/920,983** §
Filed: **August 2, 2001** §
For: **Method and System for the Automatic Segmentation of an Audio Stream into Semantic or Syntactic Units** §

Group Art Unit: **2655**
Examiner: **Huyen X. Vo**
Attorney Docket No.: **DE920000060US1**

Certificate of Transmission Under 37 C.F.R. § 1.8(a)
I hereby certify this correspondence is being transmitted via facsimile to the Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450, facsimile number (571) 273-8300 on May 30, 2006.

By: Candace Crawford
Candace Crawford

36736
PATENT TRADEMARK OFFICE
CUSTOMER NUMBER

TRANSMITTAL OF APPEAL BRIEF

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:
ENCLOSED HEREWITH:

- Appeal Brief (37 C.F.R. 41.37)

A fee of $500.00 is required for filing an Appeal Brief. Please charge this fee to IBM Corporation Deposit Account No. 09-0461. No additional fees are believed to be necessary. If, however, any additional fees are required, I authorize the Commissioner to charge these fees which may be required to IBM Corporation Deposit Account No. 09-0461. No extension of time is believed to be necessary. If, however, an extension of time is required, the extension is requested, and I authorize the Commissioner to charge any fees for this extension to IBM Corporation Deposit Account No. 09-0461.

Respectfully submitted,

Peter B. Manzo
Peter B. Manzo
Registration No. 54,700
Duke W. Yee
Registration No. 34,285
YEE & ASSOCIATES, P.C.
P.O. Box 802333
Dallas, Texas 75380
(972) 385-8777
ATTORNEYS FOR APPLICANTS

PATENT

Docket No. DE920000060US1

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of: **Haase et al.** §
§ Group Art Unit: **2655**
Serial No. **09/920,983** §
§ Examiner: **Huyen X. Vo**
Filed: **August 2, 2001** §
§
For: **Method and System for the Automatic Segmentation of an Audio Stream into Semantic or Syntactic Units** §
§

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Certificate of Transmission Under 37 C.F.R. § 1.8(a)
I hereby certify this correspondence is being transmitted via facsimile to the Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450, facsimile number (571) 273-8300 on May 30, 2006.

By: Candace Crawford
Candace Crawford

36736
PATENT TRADEMARK OFFICE
CUSTOMER NUMBER

APPEAL BRIEF (37 C.F.R. 41.37)

This brief is in furtherance of the Notice of Appeal, filed in this case on April 13, 2006.

A fee of $500.00 is required for filing an Appeal Brief. Please charge this fee to IBM Corporation Deposit Account No. 09-0461. No additional fees are believed to be necessary. If, however, any additional fees are required, I authorize the Commissioner to charge these fees which may be required to IBM Corporation Deposit Account No. 09-0461. No extension of time is believed to be necessary. If, however, an extension of time is required, the extension is requested, and I authorize the Commissioner to charge any fees for this extension to IBM Corporation Deposit Account No. 09-0461.

REAL PARTY IN INTEREST

The real party in interest in this appeal is the following party:

International Business Machines Corporation of Armonk, New York.

RELATED APPEALS AND INTERFERENCES

With respect to other appeals or interferences that will directly affect, or be directly affected by, or have a bearing on the Board's decision in the pending appeal, there are no such appeals or interferences.

STATUS OF CLAIMS

A. TOTAL NUMBER OF CLAIMS IN APPLICATION

Claims in the application are: 1, 3, 4, 6, and 8-14.

B. STATUS OF ALL THE CLAIMS IN APPLICATION

1. Claims canceled: 2, 5, and 7.
2. Claims withdrawn from consideration but not canceled: None.
3. Claims pending: 1, 3, 4, 6, and 8-14.
4. Claims allowed: None.
5. Claims rejected: 1, 3, 4, 6, and 8-14.
6. Claims objected to: None.

C. CLAIMS ON APPEAL

The claims on appeal are: 1, 3, 4, 6, and 8-14.

STATUS OF AMENDMENTS

An amendment after final rejection was not filed. Therefore, claims 1, 3, 4, 6, and 8-14 on appeal herein are as amended in the Response to Office Action dated December 22, 2005.

SUMMARY OF CLAIMED SUBJECT MATTER

A. CLAIM 1 - INDEPENDENT

The subject matter of claim 1 is directed to a method for the segmentation of an audio stream into semantic or syntactic units wherein the audio stream is provided in a digitized format (page 1, lines 4-9). A fundamental frequency is determined for the digitized audio stream (page 7, lines 2-3). Changes of the fundamental frequency are detected in the audio stream (page 7, lines 3-4). Detecting the changes of the fundamental frequency includes providing a threshold value for estimates of the fundamental frequency's voicedness and determining whether the voicedness of the fundamental frequency estimates are higher or lower than the threshold value (page 8, lines 3-9 and page 14, lines 1-3). The voicedness of the fundamental frequency estimates lower than the threshold value equals no voice and the voicedness of the fundamental frequency estimates higher than the threshold value equals voice (page 17, lines 3-6). Candidate boundaries for the semantic or syntactic units are determined depending on the detected changes of the fundamental frequency (page 7, lines 5-7). A plurality of prosodic features are extracted (page 7, lines 7-8, page 18, lines 7-9, and Figure 4C) in an environment of the audio stream where the voicedness of the fundamental frequency estimates are lower than the threshold value (page 17, lines 9-15 and Figure 4C). The environment is a period of time between 500 and 4000 milliseconds preceding and following the candidate boundaries (page 19, lines 16-20 and Figure 4C). The plurality of prosodic features are combined and boundaries for the semantic or syntactic units are determined depending only on the combined plurality of prosodic features (page 7, lines 8-10 and page 19, lines 13-15).

B. CLAIM 11 – INDEPENDENT

The subject matter of claim 11 is directed to an article of manufacture comprising a computer usable medium having computer readable program code means embodied therein for causing segmentation of an audio stream into semantic or syntactic units, wherein the audio stream is provided in a digitized format (page 1, lines 4-9), the computer readable program code means in the article of manufacture comprising computer readable program code means for causing a computer to determine a fundamental frequency for the digitized audio stream (page 7,

lines 2-3); detect changes of the fundamental frequency in the audio stream (page 7, lines 3-4), wherein detecting the changes of the fundamental frequency includes providing a threshold value for estimates of the fundamental frequency's voicedness and determining whether the voicedness of the fundamental frequency estimates are higher or lower than the threshold value (page 8, lines 3-9 and page 14, lines 1-3), and wherein the voicedness of the fundamental frequency estimates lower than the threshold value equals no voice, and wherein the voicedness of the fundamental frequency estimates higher than the threshold value equals voice (page 17, lines 3-6); determine candidate boundaries for the semantic or syntactic units depending on the detected changes of the fundamental frequency (page 7, lines 5-7); extract a plurality of prosodic features (page 7, lines 7-8, page 18, lines 7-9, and Figure 4C) in an environment of the audio stream where the voicedness of the fundamental frequency estimates are lower than the threshold value (page 17, lines 9-15 and Figure 4C), wherein the environment is a period of time between 500 and 4000 milliseconds preceding and following the candidate boundaries (page 19, lines 16-20 and Figure 4C); combine the plurality of prosodic features; and determine boundaries for the semantic or syntactic units depending only on the combined plurality of prosodic features (page 7, lines 8-10 and page 19, lines 13-15).

C. CLAIM 12 – INDEPENDENT

The subject matter of claim 12 is directed to a digital audio processing system for segmentation of a digitized audio stream into semantic or syntactic units (page 1, lines 4-9) that includes a means for determining a fundamental frequency for the digitized audio stream (page 7, lines 2-3); means for detecting changes of the fundamental frequency in the audio stream (page 7, lines 3-4), wherein detecting the changes of the fundamental frequency includes providing a threshold value for estimates of the fundamental frequency's voicedness and determining whether the voicedness of the fundamental frequency estimates are higher or lower than the threshold value (page 8, lines 3-9 and page 14, lines 1-3), and wherein the voicedness of the fundamental frequency estimates lower than the threshold value equals no voice, and wherein the voicedness of the fundamental frequency estimates higher than the threshold value equals voice (page 17, lines 3-6); a means for determining candidate boundaries for the semantic or syntactic units depending on the detected changes of the fundamental frequency (page 7, lines 5-

7); means for extracting a plurality of prosodic features (page 7, lines 7-8, page 18, lines 7-9, and Figure 4C) in an environment of the audio stream where the voicedness of the fundamental frequency estimates are lower than the threshold value (page 17, lines 9-15 and Figure 4C), wherein the environment is a period of time between 500 and 4000 milliseconds preceding and following the candidate boundaries (page 19, lines 16-20 and Figure 4C); means for combining the plurality of prosodic features; and means for determining boundaries for the semantic or syntactic units depending only on the combined plurality of prosodic features (page 7, lines 8-10 and page 19, lines 13-15).

GROUNDS OF REJECTION TO BE REVIEWED ON APPEAL

A. GROUND OF REJECTION 1 (Claims 1, 6, and 10-12)

Whether claims 1, 6, and 10-12 are unpatentable under 35 U.S.C. § 103 over Shriberg et al., *Prosody-Based Automatic Segmentation of Speech into Sentences and Topics*, Speech Communication 32 (2000) 127-154 ("Shriberg"), in view of McKiel, Jr., U.S. Patent No. 5,448,679 ("McKiel").

B. GROUND OF REJECTION 2 (Claims 3-5, 8, and 13)

Whether claims 3-5, 8, and 13 are unpatentable under 35 U.S.C. § 103 over Shriberg in view of McKiel, as applied to claim 1, and further in view of Yeldener et al., U.S. Patent No. 5,774,837 ("Yeldener").

C. GROUND OF REJECTION 3 (Claims 9 and 14)

Whether claims 9 and 14 are unpatentable under 35 U.S.C. § 103 over Shriberg in view of McKiel, in view of Yeldener, as applied to claims 8 and 13 above, and further in view of Eryilmaz, U.S. Patent No. 5,867,574 ("Eryilmaz").

ARGUMENT

A. GROUND OF REJECTION 1 (Claims 1, 6, and 10-12)

The Examiner rejects claims 1, 6, and 10-12 under 35 U.S.C. § 103 as being unpatentable over Shriberg et al., *Prosody-Based Automatic Segmentation of Speech into Sentences and Topics*, Speech Communication 32 (2000) 127-154 ("Shriberg"), in view of McKiel, Jr., U.S. Patent No. 5,448,679 ("McKiel").

The Examiner bears the burden of establishing a *prima facie* case of obviousness based on the prior art when rejecting claims under 35 U.S.C. § 103. *In re Fritch*, 972 F.2d 1260, 23 U.S.P.Q.2d 1780 (Fed. Cir. 1992). For an invention to be *prima facie* obvious, the prior art must teach or suggest all claim limitations. *In re Royka,* 490 F.2d 981, 180 U.S.P.Q. 580 (C.C.P.A. 1974). In this case, the Examiner has not met this burden because all of the features of these claims are not found in the cited references as believed by the Examiner. Therefore, the combination of Shriberg and McKiel does not reach the presently claimed invention recited in the claims.

Independent claim 1 of the present invention is representative of the claims in that group and reads as follows:

> 1. A method for the segmentation of an audio stream into semantic or syntactic units wherein the audio stream is provided in a digitized format, comprising the steps of:
>
> determining a fundamental frequency for the digitized audio stream;
>
> detecting changes of the fundamental frequency in the audio stream, wherein detecting the changes of the fundamental frequency includes providing a threshold value for estimates of the fundamental frequency's voicedness and determining whether the voicedness of the fundamental frequency estimates are higher or lower than the threshold value, and wherein the voicedness of the fundamental frequency estimates lower than the threshold value equals no voice, and wherein the voicedness of the fundamental frequency estimates higher than the threshold value equals voice;
>
> determining candidate boundaries for the semantic or syntactic units depending on the detected changes of the fundamental frequency;
>
> extracting a plurality of prosodic features in an environment of the audio stream where the voicedness of the fundamental frequency estimates are lower than the threshold value, wherein the environment is a period of time between 500 and 4000 milliseconds preceding and following the candidate boundaries;
>
> combining the plurality of prosodic features; and

determining boundaries for the semantic or syntactic units depending only on the combined plurality of prosodic features.

With regard to claim 1, the Examiner states:

> Regarding claims 1 and 11-12, Shriberg et al. disclose a method, a computer usable medium having computer readable program code, and a digital audio processing system for the segmentation of an audio stream into semantic or syntactic units wherein the audio stream is provided in a digitized format, comprising the steps of: determining a fundamental frequency for the digitized audio stream (*Section 2.1.2.3 on page 133*); detecting changes of the fundamental frequency in the audio stream (*pages 134-135, refer to figure 4*); determining candidate boundaries for the semantic or syntactic units depending on the detected changes of the fundamental frequency (*pages 134-135*); extracting a plurality of prosodic features in an environment of the audio stream where the voicedness of the fundamental frequency estimates are lower than the threshold value (*interpreted as un-voiced section, section 2.1*), wherein the environment is a period of time between 500 and 4000 milliseconds preceding and following the candidate boundaries (*section 2.1.1 on page 130, "in principle one could look at longer regions [more than 200 ms]"*); combining the plurality of prosodic features (*section 2.1.2 on page 131 and section 2.1.4 on page 137*); and determining boundaries for the semantic or syntactic units depending on the at least one prosodic feature (*pages 134-135, F0 is a prosodic feature*).
>
> Shriberg et al. fail to specifically disclose the step of detecting the changes of the fundamental frequency included providing a threshold value for estimates of the fundamental frequency's voicedness and determining whether the voicedness of the fundamental frequency estimates are higher or lower than the threshold value, and wherein the voicedness of the fundamental frequency estimates lower than the threshold value equals no voice, and wherein the voicedness of the fundamental frequency estimates higher than the threshold value equals voice. However, McKiel, Jr. teaches the step of detecting the changes of the fundamental frequency includes providing a threshold value for estimates of the fundamental frequency's voicedness and determining whether the voicedness of the fundamental frequency estimates are higher or lower than the threshold value, and wherein the voicedness of the fundamental frequency estimates lower than the threshold value equals no voice, and wherein the voicedness of the fundamental frequency estimates higher than the threshold value equals voice (*elements 26 and 30 in figure 2 and/or col. 3, line 62 to col. 4, line 6, the pitch extractor provides a pitch value which is then compared with a threshold value to determine voiced/unvoiced segment of the audio)*.
>
> Since Shriberg et al. and McKiel, Jr. are analogous are because they are from the same field of endeavors, it would have been obvious to one of ordinary shill in the art at the time of invention to modify Shriberg et al., by incorporating the teaching of McKiel, Jr. in order to enable the system to use appropriate coding

method for each of the voiced and unvoiced segments of the input speech signal to enhance coding efficiency.

Final Office Action dated March 2, 2006, pages 3-4.

Shriberg teaches prosodic modeling in controlled comparisons for speech data from two corpora: Broadcast News and Switchboard. Shriberg, page 129, section 1.4. More specifically, Shriberg teaches the motivation for each of the prosodic features and specifies the prosodic features extraction, computation, and normalization. Shriberg, page 130, section 1.4. For each inter-word boundary in Shriberg, prosodic features of the word immediately preceding and following the boundary were examined, or alternatively within a window of 20 frames or 200 milliseconds before and after the boundary. Shriberg, page 130, section 2.1.1. In other words, Shriberg teaches the use of a very specific period of time to examine and extract prosodic features in the audio stream. Thus, the only value "empirically optimized" for the method as taught by Shriberg is 200 milliseconds. Shriberg, page 130, section 2.1.1. Figure 1 of Shriberg below further illustrates this teaching:

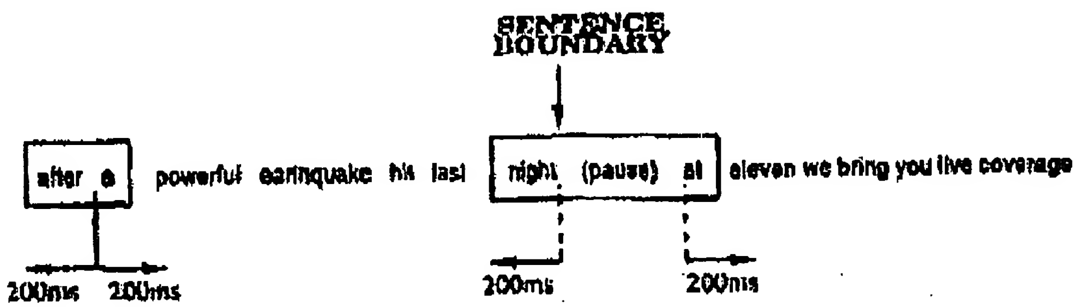

Fig. 1. Feature extraction regions for each inter-word boundary.

As Figure 1 of Shriberg clearly depicts above, the region for feature extraction is 200 milliseconds backward, or to the left, from the pause start and 200 milliseconds forward, or to the right, from the pause end.

In contrast, the present invention recites in claim 1 "extracting a plurality of prosodic features in an environment of the audio stream where the voicedness of the fundamental frequency estimates are lower than the threshold value, wherein the environment is a period of time between 500 and 4000 milliseconds preceding and following the candidate boundaries." In

other words, when the voicedness of the fundamental frequency estimates are lower than the threshold value, the audio stream contains a voiceless segment creating candidate boundaries and the plurality of prosodic features are extracted in a period of time between 500 and 4000 milliseconds preceding and following the candidate boundaries created by the voiceless segment in the audio stream. Support for this claim 1 feature may be found in the specification on page 18, line 7 – page 19, line 20 and Figure 4C. By way of example, Figure 4C of the present invention illustrates below that the plurality of prosodic features are extracted from the f1 offset candidate boundary backward, or to the left, for the exemplary period of time of 1000 milliseconds and from the f2 onset candidate boundary forward, or to the right, for another 1000 milliseconds. In addition, Figure 4C of the present invention shows that the voicedness of the fundamental frequency estimates are lower than the threshold value during the 34000-35000+ millisecond time segment.

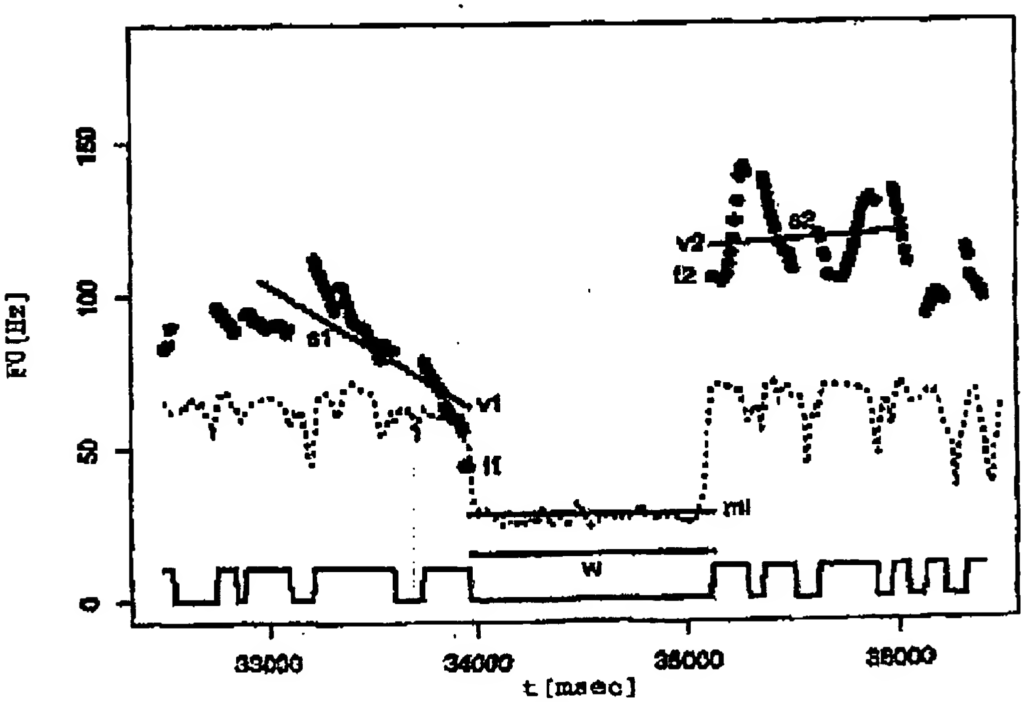

FIG. 4C

Shriberg only teaches one time value of 200 milliseconds for prosodic feature extraction, whereas, claim 1 recites a time range between 500 and 4000 milliseconds for extraction of the

plurality of prosodic features. Further, the 200 millisecond time value as taught by Shriberg is not contained within the 500-4000 millisecond time range recited in claim 1. Furthermore, it would not have been obvious to one skilled in the art, by applying the teachings of Shriberg, to provide a time range of 500-4000 milliseconds for extraction of prosodic features in order to compare results of varied time intervals for enhanced audio segmentation robustness because Shriberg only teaches the use of one specific time interval. Support for this feature may be found in the specification on page 19, lines 16-20. Therefore, Shriberg does not teach or suggest this feature recited in claim 1.

The Examiner cites Shriberg, page 130, section 2.1.1. as allegedly suggesting the use of a larger prosodic feature extraction region. Final Office Action dated March 2, 2006, page 3, lines 9-14. This Examiner-cited passage from Shriberg teaches that prosodic feature extraction was only limited to 200 ms preceding and following the boundary for practical reasons, such as "simplicity, computational constrains, and extension to other tasks," but "in principle" longer regions could be looked at. Shriberg, page 130, section 2.1.1. As the cited passage clearly indicates above, the Shriberg reference merely assumes that longer extractions regions could be used without any basis for the assumption. Shriberg was restricted to a 200 ms extraction window because of "computational constraints." In addition, even though Shriberg makes a generalized statement that "in principle one could look at longer regions," Shriberg does not specifically teach or suggest what alternative extraction periods may be used, such as, for example, 500, 1000, 2000, or 4000 ms.

Moreover, Shriberg does not teach or suggest "combining the plurality of prosodic features" as further recited in claim 1. Shriberg teaches extracting, computing, and normalizing prosodic features. Shriberg, page 130, section 1.4. Taking the Shriberg reference as a whole, the focus is "on the overall performance and on analysis of which prosodic features proved most useful for each task." Shriberg, page 130, section 1.4, lines 32-34. In other words, Shriberg determines which prosodic features are most useful by individually analyzing each prosodic feature. Shriberg discusses and analyzes each prosodic feature class in its own individual section. No section in the Shriberg reference teaches prosodic features in combination. Shriberg makes no reference to combining a plurality of prosodic features as recited in claim 1.

The Examiner cites Shriberg, page 131, section 2.1.2 and page 137, section 2.1.4. as

allegedly teaching the claim 1 feature of combining the plurality of prosodic features. Final Office Action dated March 2, 2006, page 3, lines 14-15. These Examiner-cited sections teach that a feature selection algorithm was utilized to automatically reduce the initial candidate feature set to an optimal subset. Shriberg, page 137, section 2.1.4, lines 16-18. Because the initial feature set in Shriberg contained over 100 features, the set is split into smaller subsets. Shriberg, page 137, section 2.1.4, lines 40-43. Features are grouped into broad feature classes based on the kinds of measurements involved, and the type of prosodic behavior they are designed to capture. Shriberg, page 131, section 2.1.2, lines 15-18. In other words, Shriberg only places prosodic features into groups or classes depending upon the prosodic feature's behavior or measurements and does not teach or suggest combining the plurality of prosodic features to determine boundaries for semantic and syntactic units depending only on the combined plurality of prosodic features as recited in claim 1.

Instead, Shriberg teaches "[u]sing decision tree and hidden Markov modeling techniques, to combine prosodic cues with word-based approaches" to evaluate performance. Shriberg, page 127, Abstract. In addition, Shriberg teaches that "[f]or each task...results from combining the prosodic information with language model information" was examined. Shriberg, page 130, section 1.4, lines 49-51. In other words, Shriberg teaches that prosodic information is combined with language model information to evaluate overall performance. Shriberg, page 127, Abstract. Combining prosodic and language models as taught by Shriberg is distinguishable from combining a plurality of prosodic features as recited in claim 1. Further, combining prosodic and language models to evaluate performance as taught by Shriberg is not analogous to combining a plurality of prosodic features to determine boundaries for semantic and syntactic units in an audio stream as recited in claim 1. Therefore, Shriberg does not teach or suggest this recited claim 1 feature either.

Furthermore, Shriberg does not teach or suggest "determining boundaries for semantic and syntactic units depending only on the combined plurality of prosodic features" as further recited in claim 1. Shriberg teaches that "[a]cross tasks and corpora...a significant improvement over word-only models using a probabilistic combination of prosodic and lexical information" was obtained. Shriberg, page 127, Abstract. As shown above, Shriberg teaches that inter-word boundaries are determined by the combination of prosodic and language models, whereas, claim

1 recites that semantic and syntactic unit boundaries are only determined by the combined plurality of prosodic features. As a result, Shriberg does not teach or suggest determining boundaries depending only on prosodic features as recited in claim 1 either.

McKiel does not cure the deficiencies of Shriberg. The Examiner only cites McKiel as teaching the step of "detecting changes of the fundamental frequency in the audio stream, wherein detecting the changes of the fundamental frequency includes providing a threshold value for estimates of the fundamental frequency's voicedness and determining whether the voicedness of the fundamental frequency estimates are higher or lower than the threshold value, and wherein the voicedness of the fundamental frequency estimates lower than the threshold value equals no voice, and wherein the voicedness of the fundamental frequency estimates higher than the threshold value equals voice" as recited in claim 1. Final Office Action dated March 2, 2006, page 4. McKiel makes no reference to extracting a plurality of prosodic features in an environment of the audio stream where the voicedness of the fundamental frequency estimates are lower than the threshold value, wherein the environment is a period of time between 500 and 4000 milliseconds preceding and following the candidate boundaries and combining the plurality of prosodic features to determine boundaries for the semantic or syntactic units depending only on the combined plurality of prosodic features as recited in claim 1. Consequently, McKiel does not teach or suggest these features recited in claim 1.

Since neither Shriberg nor McKiel teach or suggest extracting a plurality of prosodic features in an environment of the audio stream where the voicedness of the fundamental frequency estimates are lower than the threshold value, wherein the environment is a period of time between 500 and 4000 milliseconds preceding and following the candidate boundaries and combining the plurality of prosodic features to determine boundaries for the semantic or syntactic units depending only on the combined plurality of prosodic features as recited in claim 1, then the combination of Shriberg and McKiel cannot teach or suggest these recited claim 1 features. Therefore, the combination of Shriberg and McKiel does not teach or suggest all limitations recited in claim 1 of the present invention.

Accordingly, in view of the arguments above, Appellants respectfully urge the Board not to sustain the rejection of independent claims 1, 11, and 12 as being unpatentable over Shriberg in view of McKiel. Claims 6 and 10 are dependent claims depending on independent claim 1. As

a result, Appellants respectfully urge the Board not to sustain the rejection of dependent claims 6 and 10, at least by virtue of their dependence on independent claim 1.

B. GROUND OF REJECTION 2 (Claims 3-5, 8, and 13)

The Examiner rejects dependent claims 3-5, 8, and 13 under 35 U.S.C. § 103 as being unpatentable over Shriberg in view of McKiel, as applied to claim 1, and further in view of Yeldener et al., U.S. Patent No. 5,774,837 ("Yeldener"). This rejection is respectfully traversed. Dependent claim 5 was previously canceled in the Response to Office Action dated December 22, 2005. Consequently, the rejection of claim 5 under 35 U.S.C. § 103 is moot.

As shown in Section A above, Shriberg and McKiel, either individually or in combination, do not teach or suggest "extracting a plurality of prosodic features in an environment of the audio stream where the voicedness of the fundamental frequency estimates are lower than the threshold value, wherein the environment is a period of time between 500 and 4000 milliseconds preceding and following the candidate boundaries; combining the plurality of prosodic features; and determining boundaries for the semantic or syntactic units depending only on the combined plurality of prosodic features" as recited in independent claims 1, 11, and 12. Yeldener fails to cure the deficiencies of Shriberg and McKiel. Yeldener teaches a method for providing encoding and decoding of speech signals using voicing probability determination. Yeldener, Abstract. More specifically, Yeldener teaches:

> ...the input speech signal is represented as a sequence of time segments of predetermined length. For each input segment a determination is made as to detect the presence and estimate the frequency of the pitch F_0 of the speech signal within the time segment. Next, on the basis of the estimated pitch is determined the probability that the speech signal within the segment contains voiced speech patterns.

Yeldener, column 4, lines 25-32.

As the passage indicates above, Yeldener teaches that the prosodic feature of pitch is used to determine voiced speech pattern segments. In other words, pitch is the only prosodic feature utilized in the method taught by the Yeldener reference. Consequently, Yeldener cannot teach or suggest extracting a plurality of prosodic features, combining the plurality of prosodic features, and determining boundaries

depending only on the combined plurality of prosodic features as recited in independent claims 1, 11, and 12. Therefore, Yeldener does not teach or suggest these features recited in the independent claims.

Because neither Shriberg, McKiel, nor Yeldener teach or suggest "extracting a plurality of prosodic features in an environment of the audio stream where the voicedness of the fundamental frequency estimates are lower than the threshold value, wherein the environment is a period of time between 500 and 4000 milliseconds preceding and following the candidate boundaries; combining the plurality of prosodic features; and determining boundaries for the semantic or syntactic units depending only on the combined plurality of prosodic features" as recited in independent claims 1, 11, and 12, the combination of Shriberg, McKiel, and Yeldener cannot teach or suggest these recited independent claim features. Claims 3, 4, 8, and 13 are dependent claims depending on independent claims 1 and 12, respectively. Accordingly, Appellants respectfully urge the Board not to sustain the rejection of dependent claims 3, 4, 8, and 13 as being unpatentable over Shriberg in view of McKiel, as applied to claim 1, and further in view of Yeldener, at least by virtue of their dependence on independent claims 1 and 12.

C. GROUND OF REJECTION 3 (Claims 9 and 14)

The Examiner rejects dependent claims 9 and 14 under 35 U.S.C. § 103 as being unpatentable over Shriberg in view of McKiel in view of Yeldener, as applied to claims 8 and 13 above, and further in view of Eryilmaz, U.S. Patent No. 5,867,574 ("Eryilmaz"). This rejection is respectfully traversed.

As shown in Section B above, Shriberg, McKiel, and Yeldener, either individually or in combination, do not teach or suggest all claim limitations recited in independent claims 1, 11, and 12. In particular, Shriberg, McKiel, and Yeldener do not teach or suggest "extracting a plurality of prosodic features in an environment of the audio stream where the voicedness of the fundamental frequency estimates are lower than the threshold value, wherein the environment is a period of time between 500 and 4000 milliseconds preceding and following the candidate boundaries; combining the plurality of prosodic features; and determining boundaries for the semantic or syntactic units depending only on the combined plurality of prosodic features" as

recited in independent claims 1, 11, and 12.

These recited features above are also not taught or suggested in Eryilmaz. The Examiner only cites Eryilmaz as teaching "detecting of speech and non-speech segments comprises utilizing the signal energy or signal energy changes, respectively, in the audio stream" as recited in claim 9. Eryilmaz makes no reference to extracting a plurality of prosodic features, combining the plurality of prosodic features, and determining boundaries for the semantic or syntactic units depending only on the combined plurality of prosodic features as recited in independent claims 1, 11, and 12.

Consequently, since Shriberg, McKiel, Yeldener, and Eryilmaz, either individually or in combination, do not teach or suggest "extracting a plurality of prosodic features in an environment of the audio stream where the voicedness of the fundamental frequency estimates are lower than the threshold value, wherein the environment is a period of time between 500 and 4000 milliseconds preceding and following the candidate boundaries; combining the plurality of prosodic features; and determining boundaries for the semantic or syntactic units depending only on the combined plurality of prosodic features" as recited in amended independent claims 1, 11, and 12, then the combination of Shriberg, McKiel, Yeldener, and Eryilmaz cannot teach or suggest these features. Claims 9 and 14 are dependent claims depending on independent claims 1 and 12, respectively. Accordingly, Appellants respectfully urge the Board not to sustain the rejection of dependent claims 9 and 14 as being unpatentable over Shriberg in view of McKiel in view of Yeldener, as applied to claims 8 and 13 above, and further in view of Eryilmaz, at least by virtue of their dependence on independent claims 1 and 12.

Peter B. Manzo
Reg. No. 54,700
YEE & ASSOCIATES, P.C.
PO Box 802333
Dallas, TX 75380
(972) 385-8777

CLAIMS APPENDIX

The text of the claims involved in the appeal are:

1. A method for the segmentation of an audio stream into semantic or syntactic units wherein the audio stream is provided in a digitized format, comprising the steps of:

determining a fundamental frequency for the digitized audio stream;

detecting changes of the fundamental frequency in the audio stream, wherein detecting the changes of the fundamental frequency includes providing a threshold value for estimates of the fundamental frequency's voicedness and determining whether the voicedness of the fundamental frequency estimates are higher or lower than the threshold value, and wherein the voicedness of the fundamental frequency estimates lower than the threshold value equals no voice, and wherein the voicedness of the fundamental frequency estimates higher than the threshold value equals voice;

determining candidate boundaries for the semantic or syntactic units depending on the detected changes of the fundamental frequency;

extracting a plurality of prosodic features in an environment of the audio stream where the voicedness of the fundamental frequency estimates are lower than the threshold value, wherein the environment is a period of time between 500 and 4000 milliseconds preceding and following the candidate boundaries;

combining the plurality of prosodic features; and

determining boundaries for the semantic or syntactic units depending only on the combined plurality of prosodic features.

3. The method according to claim 1, wherein defining an index function for the fundamental frequency having a value = 0 if the voicedness of the fundamental frequency is lower than the threshold value and having a value = 1 if the voicedness of the fundamental frequency is higher than the threshold value.

4. The method according to claim 3, wherein extracting the plurality of prosodic features is in an environment of the audio stream where the value of the index function is = 0.

6. The method according to claim 1, wherein at least one prosodic feature is represented by the fundamental frequency.

8. The method according to claim 1, further comprising first detecting speech and non-speech segments in the digitized audio stream and performing the steps of claim 1 thereafter only for detected speech segments.

9. The method according to claim 8, wherein the detecting of speech and non-speech segments comprises utilizing the signal energy or signal energy changes, respectively, in the audio stream.

10. The method according to claim 1, further comprising the step of performing a prosodic feature classification based on a predetermined classification tree.

11. An article of manufacture comprising a computer usable medium having computer readable program code means embodied therein for causing segmentation of an audio stream into semantic or syntactic units, wherein the audio stream is provided in a digitized format, the computer readable program code means in the article of manufacture comprising computer readable program code means for causing a computer to effect:

determining a fundamental frequency for the digitized audio stream;

detecting changes of the fundamental frequency in the audio stream, wherein detecting the changes of the fundamental frequency includes providing a threshold value for estimates of the fundamental frequency's voicedness and determining whether the voicedness of the fundamental frequency estimates are higher or lower than the threshold value, and wherein the voicedness of the fundamental frequency estimates lower than the threshold value equals no voice, and wherein the voicedness of the fundamental frequency estimates higher than the threshold value equals voice;

determining candidate boundaries for the semantic or syntactic units depending on the detected changes of the fundamental frequency;

extracting a plurality of prosodic features in an environment of the audio stream where the voicedness of the fundamental frequency estimates are lower than the threshold value, wherein the environment is a period of time between 500 and 4000 milliseconds preceding and following the candidate boundaries;

combining the plurality of prosodic features; and

determining boundaries for the semantic or syntactic units depending only on the combined plurality of prosodic features.

12. A digital audio processing system for segmentation of a digitized audio stream into semantic or syntactic units comprising:

means for determining a fundamental frequency for the digitized audio stream;

means for detecting changes of the fundamental frequency in the audio stream, wherein detecting the changes of the fundamental frequency includes providing a threshold value for estimates of the fundamental frequency's voicedness and determining whether the voicedness of the fundamental frequency estimates are higher or lower than the threshold value, and wherein the voicedness of the fundamental frequency estimates lower than the threshold value equals no voice, and wherein the voicedness of the fundamental frequency estimates higher than the threshold value equals voice;

means for determining candidate boundaries for the semantic or syntactic units depending on the detected changes of the fundamental frequency;

means for extracting a plurality of prosodic features in an environment of the audio stream where the voicedness of the fundamental frequency estimates are lower than the threshold value, wherein the environment is a period of time between 500 and 4000 milliseconds preceding and following the candidate boundaries;

means for combining the plurality of prosodic features; and

means for determining boundaries for the semantic or syntactic units depending only on the combined plurality of prosodic features.

13. An audio processing system according to claim 12, further comprising means for generating an index function for the voicedness of the fundamental frequency having a value = 0 if the voicedness of the fundamental frequency is lower than a predetermined threshold value

and having a value = 1 if the voicedness fundamental frequency is higher than the threshold value.

14. Audio processing system according to claim 12 or 13, further comprising means for detecting speech and non-speech segments in the digitized audio stream, particularly for detecting and analyzing the signal energy or signal energy changes, respectively, in the audio stream.

EVIDENCE APPENDIX

There is no evidence to be presented.

RELATED PROCEEDINGS APPENDIX

There are no related proceedings.